Functional profiling of the sequence stockpile: a review and assessment of *in silico* prediction tools

Prabakaran R[1,2,3*], Bromberg Y[1,3,4*]

1 Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ 08873, USA
2 Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ 08901, USA
3 Department of Biology, Emory University, Atlanta, GA 30322, USA
4 Department of Computer Science, Emory University, Atlanta, GA 30322, USA
* Corresponding author: yana.bromberg@emory.edu, prabakaran@emory.edu

Abstract

In silico functional annotation of proteins is crucial to narrowing the sequencing-accelerated gap in our understanding of protein activities. Numerous function annotation methods exist, and their ranks have been growing, particularly so with the recent deep learning-based developments. However, it is unclear if these tools are truly predictive as most do not identify new terms in functional ontologies. We thus ask, can they identify sequences capable of carrying out known molecular functions but are non-homologous to or far-removed from known protein families? Here, we explore the potential and limitations of the existing methods in predicting molecular functions of thousands of such orphan proteins. Lacking the ground truth functional annotations, we translated the assessment of the function prediction into evaluation of functional similarity of orphan siblings. Notably, our approach transcends the limitations of functional annotation vocabularies and provides a platform to compare different methods without the need for mapping terms across ontologies. We find that most existing methods are limited to homology-based annotations and are thus descriptive, rather than predictive of function. Curiously, despite their seemingly unlimited by homology scope, novel deep learning methods remain far from capturing functional signal encoded in protein sequence. We believe that our work will inspire the development of a new generation of methods that push our knowledge boundaries and promote exploration and discovery in the molecular function domain.

Key words: Protein function prediction, annotation, Ontology, Homology, Deep-learning, Assessment

Introduction

A typical cell contains about 0.2 g/ml proteins, which translates to up to a billion molecules per cell[1, 2]. However, the corresponding number of distinct protein sequences varies from only a few hundred to tens of thousands. Characterizing these vital biomolecular nanomachines, i.e. identifying their cellular functions, associated pathways, localization, interaction partners, and catalytic activities, is crucial for understanding their role in cellular biology. Experimental annotation of protein function remains significantly limited by its cost and speed. For example, among the 94.5 million protein sequences that have been deposited in UniProt in the last three years, only 6,974 (<0.01%) were manually curated. Thus, the growing influx of sequencing data has necessitated accurate computational annotation of protein function for diverse downstream analyses.

Over the last two decades, the number of bioinformatics tools developed for *in silico* protein annotation has grown and algorithms diversified. Historically, the most common and reliable techniques for annotation relied on the transfer of function by homology, i.e. shared ancestry resulting in sequence similarity. To characterize a given query protein, various alignment and domain profiling tools such as BLAST, PSI-BLAST, and HMMER[3-7] were used to search annotated protein databases[8-11]. More recently, faster algorithms have been developed to process and annotate large sequence datasets, including sequence reads and proteins extracted from genome assemblies[12-16]. The challenges associated with protein functional annotation are multi-fold and have been discussed earlier[17-20]. To summarize, there are three bottlenecks in accurate functional annotations: evolutionary caveats, lack of experimental annotations, and limitations of functional ontologies.

The first two sets of challenges pertain to the difficulty of identifying the levels of homology sufficient to transfer existing annotations to new sequences and, further, to the overall dearth of experimental annotations. As life evolves and adapts, divergent evolutionary processes result in homologous genes of different functions, which could further end up as false positive functional annotations of sequence and structurally similar proteins[21]. One such example is the enzymatically inactive duck δ crystallin I

that shares >90% sequence identity with an active δ crystallin II[21, 22]. At the same time, different genes converging to perform same function may have minimal homology – a false negative[23, 24]. For example, human (PDB:1PL8) and *Rhodobacter sphaeroides* (PDB:1K2W) sorbitol dehydrogenases are sequence different.

To further complicate the issue, diverged genes found in different species, i.e. orthologs, that do participate in the same molecular mechanisms, may not operate at the same rate or efficiency given the specific species' environmental constraints – a functional difference that is often ignored. Annotations are further biased towards common proteins from large families and to species of interest. For example, experimental evidence for GO annotations only exists for less than 15% of proteins in SwissProt. The effects of these biases are compounded by the computational annotation of newly accumulated genomic data – a process that fosters annotation error propagation. Note that the existing functional annotations can, by default, only cover the observed part of the protein universe, i.e. annotation of new sequences may be flawed simply by our limited knowledge of biotic functional capacity (**Figure 1**). In short, the classical approach of transferring protein function by homology is complicated by convergent/divergent evolution, lack of annotation coverage, and errors in existing annotations; it is also limited to existing classes of proteins, reducing chances of discovery of novel functions.

The third challenge is more technical in nature. The task of representing the ambiguous, environment-dependent, hierarchical role of a given protein with a set of human-understandable ontology terms is exceedingly difficult[11, 25-31]. Depending on the level of granularity and environmental conditions, a protein's function could vary widely. For example, all kinases are phosphotransferases that catalyze the transfer of phosphate from ATP to carbohydrates, lipids, or proteins. However, kinases are part of almost every cellular process and many metabolic pathways, i.e. they can be assigned a wide range of functions. On the other hand, proteins involved in the same biological pathway have different catalytic functions almost by definition; e.g. glycolysis (map00010;[26, 32]) involves kinases and dehydrogenases. That is, different molecular functions can contribute to the same biological role, while proteins of the same

molecular function may have different biological roles. An ideal protein function ontology should be robust to this variability, but also precise, widely applicable, expandable, and, lately, machine-readable. This ontology does not yet exist.

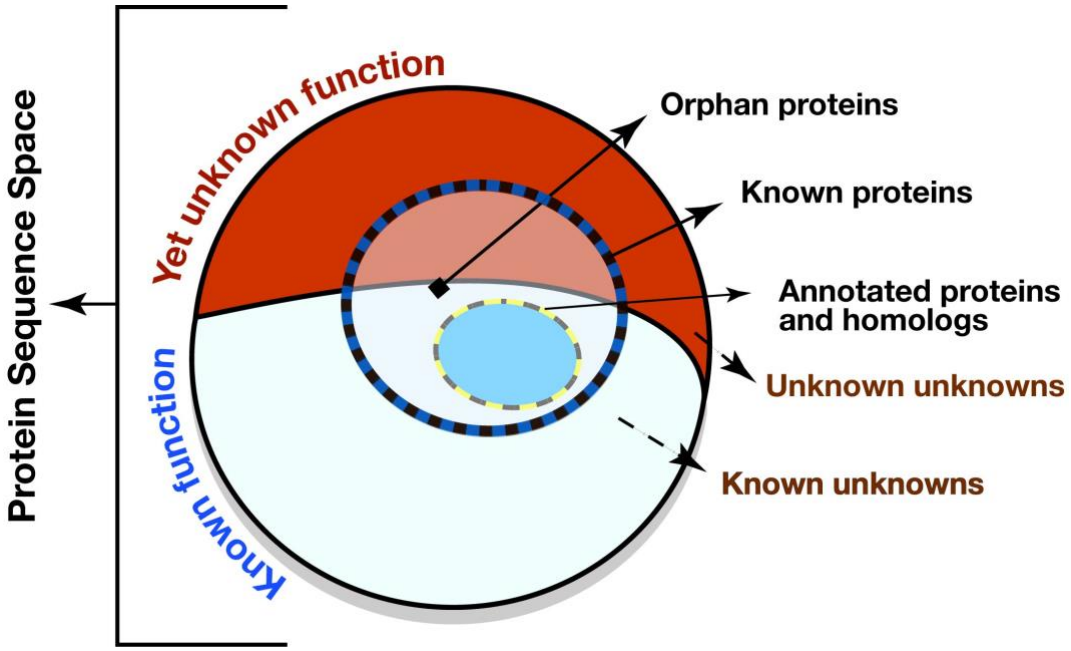

Figure 1: The limits of protein function annotations. Of the complete set of proteins (entire circle), containing known/previously observed proteins (blue/black dashed circle outline) and unknown/not-yet seen proteins, some fraction carries out unknown functions (fraction of circle in red) rather than known (fraction in white) ones. Existing experimental (and homology-based) protein function annotations (blue oval) cover a small part of the complete protein sequence space. The stockpile of orphan proteins, i.e. those lacking annotation and having no known homologs, is growing as we explore previously untouched environments with better and faster gene/protein capture tools**.** This manuscript is focused on evaluating function prediction method performance for this orphan set.

A significant amount of research has gone into targeting these challenges to computational function prediction. Critical Assessment of Functional Annotation algorithms (CAFA) is a community experiment that provides an even ground for the assessment of existing methods[33, 34]. CAFA employs a time-delayed evaluation where predictions of functions of a large set of yet to be annotated genes/proteins are collected and assessed over a period of time through wet-lab experiments. CAFA results have documented the continuous emergence of new, better-performing prediction methods. Research has moved beyond sequence comparison, introducing

new computational techniques, and incorporating additional biological data such as the protein-protein interactions, expression, phenotypic changes due to mutation, etc.

A key point in recent method development is the representation of protein sequences as embeddings, i.e. projections of proteins into the latent space. Embeddings are interpretations of deep neural networks, learnt in the process of addressing a predefined objective function[35, 36]. Protein sequence embeddings have been successful in annotating various protein features, but most obviously protein structure[37-39]. Recently, deep learning methods have been developed to annotate protein function. For example, Littmann et al. have explored the application of protein embeddings to function annotation, reporting performance on par with CAFA ‘s top 10 best performers[40, 41]. However, besides embeddings capturing protein structural aspects and thus informing function, it remains unclear exactly which (or whether) aspects of functionality are captured by embeddings.

One important inference from CAFA’s experience is the challenge of establishing metrics for the assessment of methods. That is, what is to be considered a correct annotation for a given protein? Given a protein P_1 with that carries out functions f_1, f_2, and f_3 defined by a relevant ontology, if a method M_1 predicts the protein to be of function f_1 only, is this a correct annotation? How does this method perform in comparison to M2, which predicts the protein to carry out f_3, f_4, and f_5? While for an individual annotation, say f_1 *vs.* f_2, ontology distance metrics can be established[42-45], evaluating multiple annotations per protein is harder. Adding to the problem is the incomplete “ground truth” annotation, i.e. how would one take into account the protein’s unknown cellular roles? Here, we provide a method and ontology-blind assessment approach for comparison of protein annotation tools. We evaluate the predictions of computational methods for a set of proteins, sharing little sequence similarity with proteins in available databases. We ask, what is a correct annotation for a protein with no known homologs? To answer this question, we use structural similarity and deep learning-based techniques to establish whether a protein pair in our set shares functionality, regardless of what specifically the protein does. We then evaluate other methods’ ability to recall shared functions for these pairs.

METHODS

Extracting the test data set. From the ESM Metagenomic Atlas (proteins translated from metagenome records of the MGnify database), we collected 53.5M (53,501,759) protein sequences, translated from metagenome-assembled genes, and having high-confidence predicted 3D structures, i.e. predicted local distance difference test (pLDDT) and predicted TM-score (pTM) greater than 0.9[37, 38, 46-48]. Note that our selected sequences make up less than a tenth of all structures in ESM and represent a significantly smaller fraction still of all metagenome-derived proteins MGnify collected over the years. Thus, our evaluation is limited to a subset of ordered proteins, whose structure is well predictable and, thus, essentially like that of experimentally studied proteins.

These sequences were aligned against UniRef100 (mmseqs2 at default sensitivity =5.7[49, 50]). Sequences in UniRef are often used as reference for function transfer and as the training set for prediction models[9, 49]. To avoid using training data in method testing, we focused on "orphan" sequences, i.e. those that do not share homology with proteins in UniRef. We identified 54,359 orphans with less than 30% sequence identity to UniRef100. To further simplify the evaluation task, we filtered out longer proteins (over 400 residues) that are likely to contain multiple domains, as well as proteins whose sequences are truncated in the corresponding MGnify contigs, to retain 11,484 sequences. To avoid excess focus on sequence similarity, we further sequence-reduced this set at 90% identity using CD-HIT[51]. The final dataset of orphan proteins contained 11,444 proteins with ESM predicted structures and corresponding MGnify cDNA sequences.

We coupled these 11.4K orphan proteins into functionally similar pairs, "orphan siblings," through a combination two orthogonal approaches: structural similarity (via TM-align) and function similarity (via the Siamese Neural Network, SNN[52]). In our earlier work, this approach attained 90% precision in identifying enzymes of the same function (EC number, at TM-score=0.7 and SNN-score = 0.98)[52]. We ran Foldseek[53] (with TM threshold=0) to compare the structures of all 11.4K proteins in our set amongst themselves. Only ~309K of these protein pairs (0.5% of ~65M possible ones)

were deemed structurally similar enough to be aligned. We further annotated functional similarity scores for these 309K pairs via SNN (**Figure 2**). Only 6K (6,219, 2% of 309K) pairs attained the pre-set cutoffs for shared similarity (TM score≥0.7, SNN score≥0.98) and were thus labelled *orphan siblings* making up our *test set*. Note that at these cutoffs, proteins that are not identified as being of the same function can still be functionally identical (recall = 30%). Thus, we primarily assessed the performance of protein function annotation tools based on their recall, i.e. the capacity to find all test set orphan siblings, rather than on their precision of labelling pairs as functionally similar. Note that as our method for identifying shared function is also computational and subject to thresholding, we further explored the performance of the selected prediction methods by varying the TM score and SNN score cut-offs (**Figures 3, 4** and ***SI Table 1***).

Translating predicted annotations into functional similarity. We reformulated the function prediction challenge to overcome the limitations of evaluating and comparing methods that target different functional vocabularies; i.e. tools predicting Gene Ontology (GO) molecular function (MF) terms or Enzyme Commission (EC) numbers. We also included tools that identify protein Pfam domains and assign sequences to KO and COG ortholog groups[11, 54, 55]. Note that while GO and EC aim to explicitly describe the function of the protein, Pfam and Ortholog methods capture protein families and evolutionary relationships, which are related to, but not necessary directly reflective of function. If a protein was annotated with multiple Pfam domains, EC numbers, or orthologous groups, we retained all annotation labels. We selected 13 protein annotation tools for our assessment based on the availability of a standalone version or a web server that can process multiple sequences. These 13 methods are grouped based on the type of protein annotation into four categories: GO – DeepFri[56], DeepGOPlus[57], GoPredSim[41], GOProFormer[58], and NetGO[59], Pfam – HMMER[60], InterProScan[61], ProtENN & ProtCNN[62] and ProteInfer[63], Orthologs – GhostKOALA[64], KofamKOALA[65] and E.C. – ECPred[66] and Mantis[67]. Methods such as Mantis[6, 67] and ProteInfer[63] were included in more than one category since they provide multiple protein annotations (**Table 1**). We also evaluated the value of using pairwise

DNA/Protein embedding similarities from unsupervised models such as Bepler[68], CPCProt[69], ESM-2[38], LookingGlass Encoder[70], ProtTrans[36], SeqVec[71] and Word2Vec[72]. In addition, we used SwiftOrtho[73], a method that identifies orthologous pairs in a given set of proteins[74, 75], as an upper bound of homology-based evaluation of our test set.

To quantify method prediction accuracy, we scored the similarity between annotations of proteins in each pair. Consider a protein pair P_1 and P_2 predicted by method M_1 to carry out sets of functions Fu_1 and Fu_2, respectively. Fu_1 (and Fu_2) consist of several annotation terms from GO, EC, Pfam, or ortholog groups as assigned by M_1; each term is assigned a prediction score (or E-value) and is accepted or rejected at a prediction score threshold (τ). The similarity (S) between F_1 and F_2 is defined by the Jaccard similarity coefficient, i.e. the ratio of the intersection set of terms to the union set (**Eqn. 1**). In case of GO annotations, we used the information accretion term[42, 76] (I_a, **Eqn. 2 and 3**) to weigh the GO term assignment according to term frequency of appearance among UniProt GO annotations with experimental evidence[77, 78]; information accretion of a child GO term v, $I_a(v)$, is the information gained by adding v to its parent term(s).

$$S(P1(\tau),\ P2(\tau)) = \frac{|Fu1 \cap Fu2|}{|Fu1 \cup Fu2|} \qquad \text{(Eqn. 1)}$$

$$S_{GO}(P1(\tau),\ P2(\tau)) = \frac{\sum_{v \in Fu1 \cap Fu2} I_a(v)}{\sum_{v \in Fu1 \cup Fu2} I_a(v)} \qquad \text{(Eqn. 2)}$$

$$I_a = -\log\left(\Pr(v \mid Parent(v)) \right) \qquad \text{(Eqn. 3)}$$

Additionally, the similarity score for a protein pair (P_1 and P_2) given their embeddings (E_1 and E_2) of length l, was derived from the Euclidean and cosine distances (**Eqn. 4 and 5**).

$$S_{Euclidean}(P1,\ P2) = \frac{0.5}{0.5 + \sqrt[2]{\sum_{i=1}^{l} ({E_1}^i - {E_2}^i)^2}} \qquad \text{(Eqn. 4)}$$

$$S_{Cosine}(P1,\ P2) = \frac{E_1 . E_2 + 1}{2} \qquad \text{(Eqn. 5)}$$

Assessment of prediction performance. The performance of a given method in identifying orphan siblings was measured by computing the difference in similarity

scores (ΔS) between orphan sibling and non-sibling pairs (**Eqn. 6**). We weighted the similarity difference by Recall values to penalize methods for failing to identify our gold standard sibling pairs. ΔS varies in range [-1,1], where a positive value indicates higher similarity scores for sibling vs. non-sibling pairs.

$$\Delta S(\tau) = Recall * \left(\langle S_{(P_1, P_2) \epsilon sibings}(P_1, P_2) \rangle - \langle S_{(P_1, P_2) \epsilon non-sibings}(P_1, P_2) \rangle \right) \quad \textbf{(Eqn. 6)}$$

We additionally computed the Area under Precision-Recall curve (PR AUC), area under Receiver Operating Characteristic curve (ROC-AUC), and $F1_{max}$ by varying the similarity score threshold (τ^S) for calling a protein pair functionally identical. Note that for these calculations, to remove the imbalance in the number of sibling (positives) vs. non-sibling (negative) pairs, we undersampled the non-siblings pairs set to match the number of sibling pairs; we repeated the under-sampling 20 times and computed the average and standard deviation of all measures. Finally, we chose the optimum prediction score threshold (τ_{opt}) and a similarity score threshold ($\tau^S{}_{opt}$) corresponding to the maximum $F1_{max}$ and computed Precision and Recall (**Eqn. 7-9**) for all methods.

$$Recall = \frac{\#\ of\ correctly\ identified\ siblings}{\#\ of\ acutal\ siblings\ in\ the\ dataset} \quad (\mathbf{Eqn.\ 7})$$

$$Precision = \frac{\#\ of\ correctly\ identified\ siblings}{\#\ of\ predicted\ siblings} \quad \textbf{(Eqn. 8)}$$

$$F_1\ score = \frac{2\ (Precision \times Recall)}{Precision + Recall} \quad \textbf{(Eqn. 9)}$$

The F_1 score, as the harmonic mean of Recall and Precision, provides an overall prediction performance in identifying siblings. However, consider that our gold standard set is based on SNN score and TM score measures that, when taken together, only achieved 30% Recall in testing, i.e. bearing 70% of false negatives[52]. As such, in this work, precision and F_1 measures, which are concerned with the number of putative false positives, may have unfairly penalized method performance. To even out estimates of performance, we thus also evaluated methods at thresholds corresponding to 90% Recall. Further, the statistical significance of the performance differences between the tools in terms of ΔS, F_{1max}, PR-AUC and ROC-AUC were assessed through the Wilcoxon rank-sum test and Student's t-test using SciPy[79].

RESULTS AND DISCUSSION

No one method is best for function annotation. From metagenomic data, we collected a set of proteins having no homology to any of the UniRef100 sequences (*orphans*) and paired them by expected functional identity (*orphan siblings*, Methods; **Figure 2**). Note that our definition of orphan siblings is not dependent on protein sequence similarity. Among the 309k orphan protein pairs, 99.3% (307,434) shared <40% sequence identity, while 0.08% (240) were ≥80% identical (**Figure 2**). TM and SNN cutoffs of 0.7 and 0.98[52], respectively, labelled 6,219 of these protein pairs as siblings; among them 5,576 (89.7%) pairs had less than 40% identity and 95 (1.53%) pairs shared ≥80% sequence identity, i.e. a slight enrichment for sequence similarity among orphan siblings as compared to the complete set of orphans. Despite this enrichment, siblings were largely composed of sequence dissimilar protein pairs.

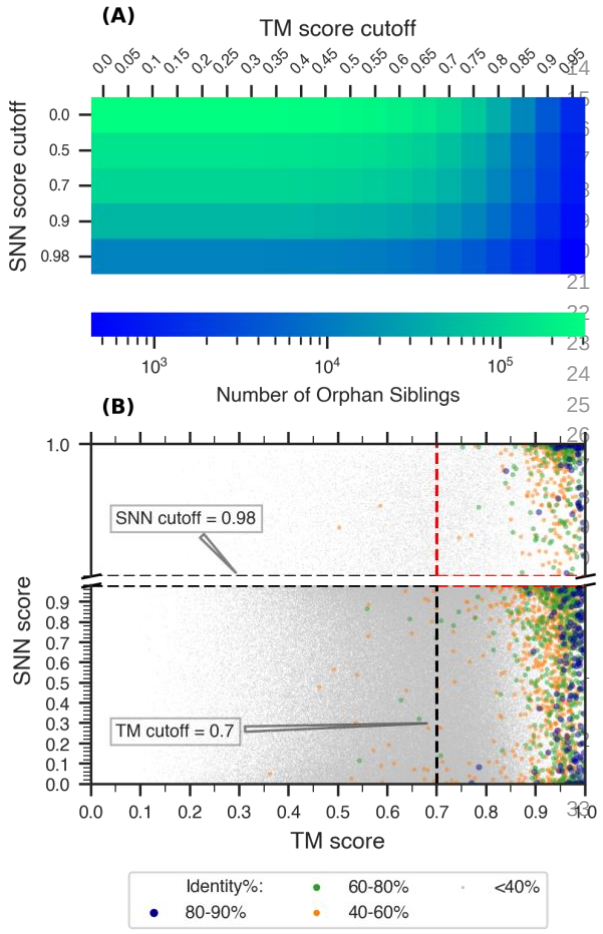

Figure 2: Defining the Orphan sibling dataset. Orphan protein pairs with similar functionality are termed *orphan siblings*. Orphan siblings are identified using the structural alignment TM score (x-axis) and the SNN functional similarity score (y-axis; Methods). **(A)** The number of orphan pairs considered siblings at each position on the plot is highlighted according to a log-scale gradient scheme from few proteins (blue) to many proteins (green). In **(B)** each dot represents a protein-pair and is colored by sequence identity from very low (< 40%; gray) to very high (80-90%; dark blue). Note that no pairs over 90% identity were included in our set. The dashed lines indicate the TM and SNN score cut-offs (0.7 and 0.98) chosen for this work.

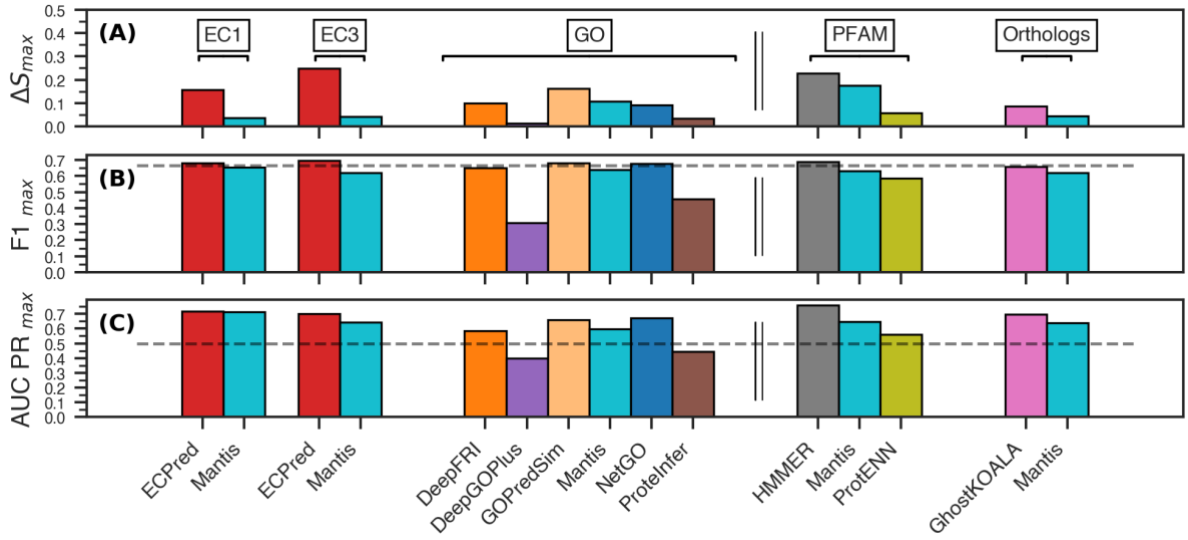

Figure 3: Comparing method performance in each annotation category. The bar plots show the highest method **(A)** ΔSimilarity scores, **(B)** F1 scores, and **(C)** Area under the Precision-Recall curves for the prediction of orphan siblings (TM-score and SNN cutoffs of 0.7 and 0.98, respectively), based on: Enzyme Commission numbers (EC1 – first digit, EC3 – up to third digit), Gene Ontology Molecular Function terms, Pfam domains, and Ortholog groups. The highest scores were selected for each method by varying the prediction score threshold. The double line separates explicit functional annotations (EC and GO) from Pfam and Ortholog definitions. The gray dashed lines indicate the performance of the random classifier.

We measured the ability of existing functional annotation methods to assign identical functions to each protein in the orphan sibling pair (**Figure 3**, **Table 1).** Notably, the top performers in this evaluation were not restricted to any one class of annotation; i.e. ECPred, GOPredSim, Pfam HMMER, and GhostKOALA attained similar performance ($F1_{max}$, ΔS, etc.; **Table 1**).

Curiously, except for GOPredSim, which uses ProtT5 embeddings[36], the deep-learning models built to predict GO terms did not top the list. While NetGO attained solid performance, though still lower than the best model, other deep-learning methods (DeepFRI, DeepGOPlus, GOProFormer, and ProteInfer) were significantly worse (Wilcoxon rank-sum test, all p-values < $1E^{-8}$). Note that GOProFormer was developed on yeast proteins[58] and using it to predict microbial protein function may have been beyond its scope of work.

Protein embedding distances are not directly informative of functional similarity. We further computed similarity between protein pairs based on cosine and Euclidean

distance between embeddings vectors of all 309K pairs. Surprisingly, none of these distances captured the functional similarity between protein pairs well. Among the seven different embedding methods that were assessed, only ESM-2, ProtT5 and LookingGlass showed somewhat higher performance than a random classifier (**Figure S1**). Note that the generic choice of architecture was not the primary driving factor of performance — among the top three embeddings, ESM-2 and ProtT5 are transformer-based protein language models[36, 38, 80] whereas LookingGlass is a bi-directional LSTM model trained on short DNA reads[70]. The other four embeddings were SeqVec (LSTM) and Word2Vec (neural network) inspired by Natural Language Processing[71, 72] , Bepler is also a bi-directional LSTM trained on amino acid sequences[68] and CPCProt is a convolution encoder trained through contrastive learning to identify subsequent fragments of protein against random protein fragments[69].

We also note that given that ESM embeddings could predict protein structures, we expected these to capture functional similarity as defined, in part, by structural alignments. However, interpretation of language models is complicated and the extraction of average representation from multiple layers or a representation of any particular layer is bound to lead to loss of information[81, 82]. Our results thus highlight limitations of cross-domain application of unsupervised deep learning models without extensive analysis and fine-tuning.

Function prediction strongly linked to protein structure. Our definition of siblings is dictated by structural similarity (TM alignment) and functional similarity (SNN). By using a high SNN cut-off (0.98), we have negotiated significant reduction in false positives at a loss of true positives. In other words, our approach to identifying functional siblings, while being very accurate (90% precision), is known to miss 70% of protein pairs annotated to be of the same function (30% recall). While our structural similarity cutoff of TM score≥0.7 is an accepted value[83] and the SNN cut-off of score≥0.98 was confirmed by our earlier experiments[52], we aimed to explore method behaviour across the complete range of protein similarities. We thus computed the prediction performance of methods by varying the TM (**Figure 4**) and SNN score (**Figures S2, S3**) cut-offs and, thus, redefining the protein pairs considered siblings.

Note that in evaluating structural similarity we focused on pairs identified by Foldseek as possibly alignable (Methods), i.e. 309K protein pairs of 65M possibilities, but for these we varied the TM-score in the [0,1] range. The SNN method was trained to recognize pairs as functionally similar above the 0.5 cutoff, so we explored the SNN scores in the [0.5,1] range.

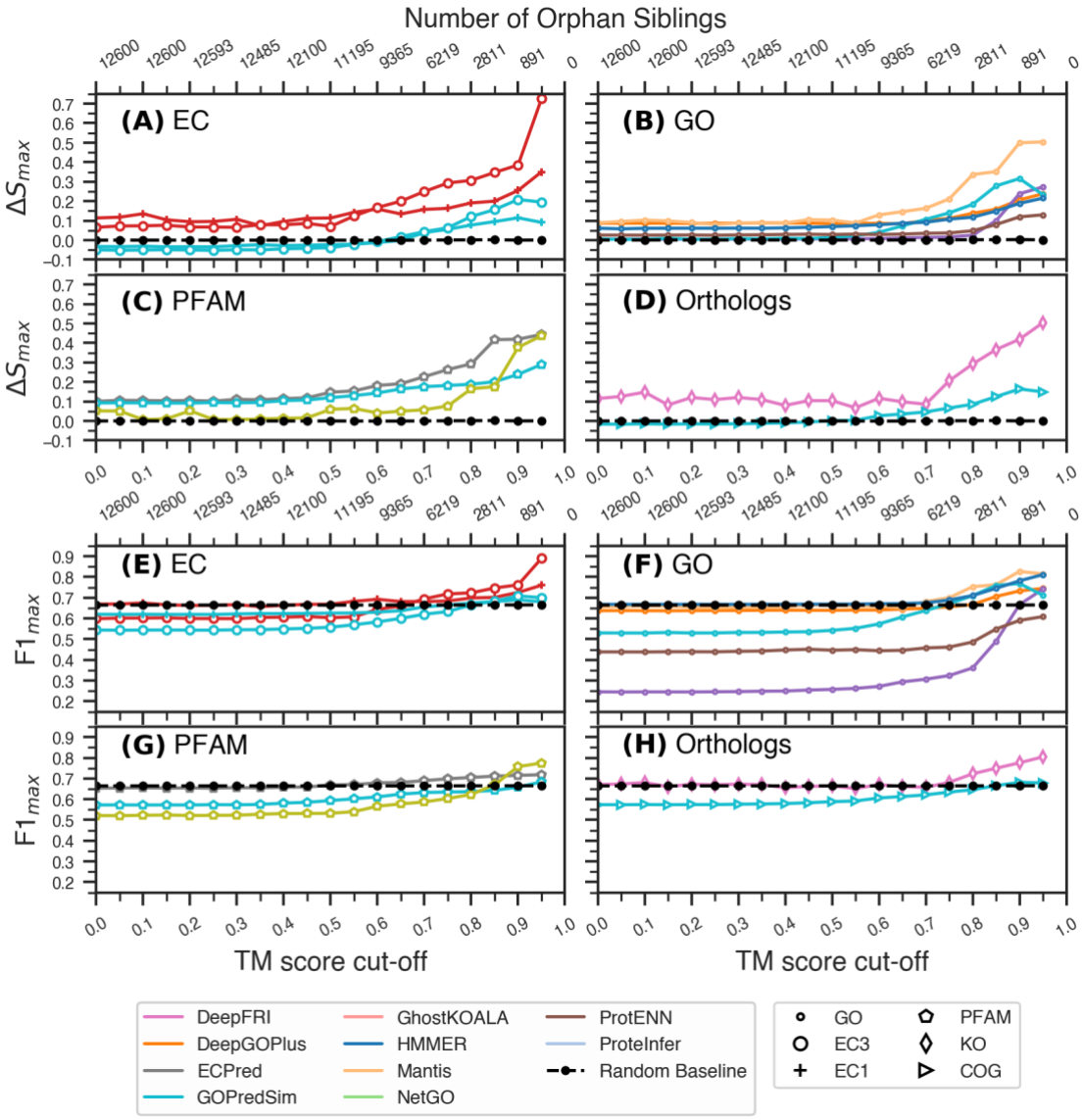

Figure 4: Variations in performance over TM-score cut-offs. The highest **(A-D)** ΔSimilarity and **(E-H)** $F1_{max}$ scores (y-axis) of top performing methods vary based on predicted protein annotations. All scores were computed at SNN cutoff =0.98 and at different TM-score cutoffs (bottom x-axis), resulting in a different number of orphan sibling pairs (top x-axis). The average performance of 100 random baseline classifiers is plotted for comparison in each panel (dotted black line).

The top-performing methods (ECPred, GOPredSim, HMMER and GhostKOALA) showed consistently better performance than other methods across different TM and SNN score cut-offs. As expected, we observed a steep rise in performance of all methods at TM-score≥0.7 confirming that structural similarity above that threshold plays a significant role in functional similarity. At the same time, restricting the SNN score cutoff to 0.98 increased the performance (AUC under Precision-Recall curve and F1-score) of ECPred, ProteInfer, ProtCNN and DeepGOPlus by at least 20% (**Figure S2**) but reduced the performance of GhostKOALA. SNN captures functional similarity independent of either sequence or structural similarity[52]. We thus expect that tightening the SNN threshold reduces the performance of methods with strong dependency on sequence similarity, such as GhostKOALA.

What do the best performing methods capture? To answer this question, we evaluated contributions of known functionally relevant factors, i.e. sequence and structure, to method functional annotations.

We first clustered our orphan proteins, by sequence identity at different cut-offs between 0.9 to 0.4 using CD-HIT[51] and explored their shared functionality (**Table 2, S1, S2 & Figure 5**). We found that less than a tenth of a percent of protein pairs in our orphan dataset set (240 pairs) were highly sequence-similar (80-90% seq.id), while the vast majority were not; i.e. more than 99% of protein pairs shared less than 40% sequence identity. Among all method predictions we observed significant enrichment of sequence similar protein pairs (≥40% seq. id.) and a depletion of dissimilar pairs. Note that this observation is unrelated to the putative correctness of their functional annotations.

We further observed that the enrichment of GOPredSim and ECPred predictions was limited only to pairs of higher similarity (≥60%). GhostKOALA's predictions, on the other hand, were not significantly enriched in highly sequence similar pairs (80-90%). Note that the small number of these highly sequence similar pairs complicates inference. That is, of the 240 such pairs, GOPredSim and ECPred identified ten and two, respectively – a small, but significant number – while GhostKOALA, which is built to annotate proteins using ortholog information, identified only one. We expect that the

latter result is due to our test set being made up of orphan proteins, i.e. those without homologs in the predictor's reference database. These observations suggest that, as expected, sequence information is important in driving function annotation by all methods, but GOPredSim and ECPred are more reliant than others on high sequence similarity.

The enrichment in the number of siblings predicted by each method within sequence-similar bins did not correlate with higher function prediction accuracy across these bins. For example, ECPred precision was worse for high similarity pairs than for lower ones, i.e. the opposite of the enrichment trend (**Table 2 and S2**). On the other hand, GOPredSim precision was similar for all sequence identity bins. HMMER which showed significant enrichment in the 40-80% sequence identity bin, had the highest precision of 63% in the 80-90% identity space. To summarize, while the predictions of the methods in this study are biased towards identifying sequence similar proteins as siblings, the accuracy of such predictions in terms of the functional similarity of the two does not agree with this assessment.

We also explored this homology-based functional annotations using SwiftOrtho[73] – a method that identifies orthologous pairs in a given set of proteins based on sequence similarity[74, 75]. As expected, SwiftOrtho correctly identified 239 of 240 of the highly-sequence similar (80-90% sequence identity) pairs; it's predictions were also enriched in pairs of sequence similar proteins at all levels of similarity ≥40% (**Table 2 & S1**). These results highlight the success achievable by homology-based methods in the presence of the relevant reference sets and further emphasize their deficiency in the absence of such reference. We note, however, that distinguishing orthologs from paralogs is hard[84, 85] and even harder without the taxonomic and/or genomic context. In fact, SwiftOrtho also (putatively) incorrectly labelled 144 protein pairs as siblings.

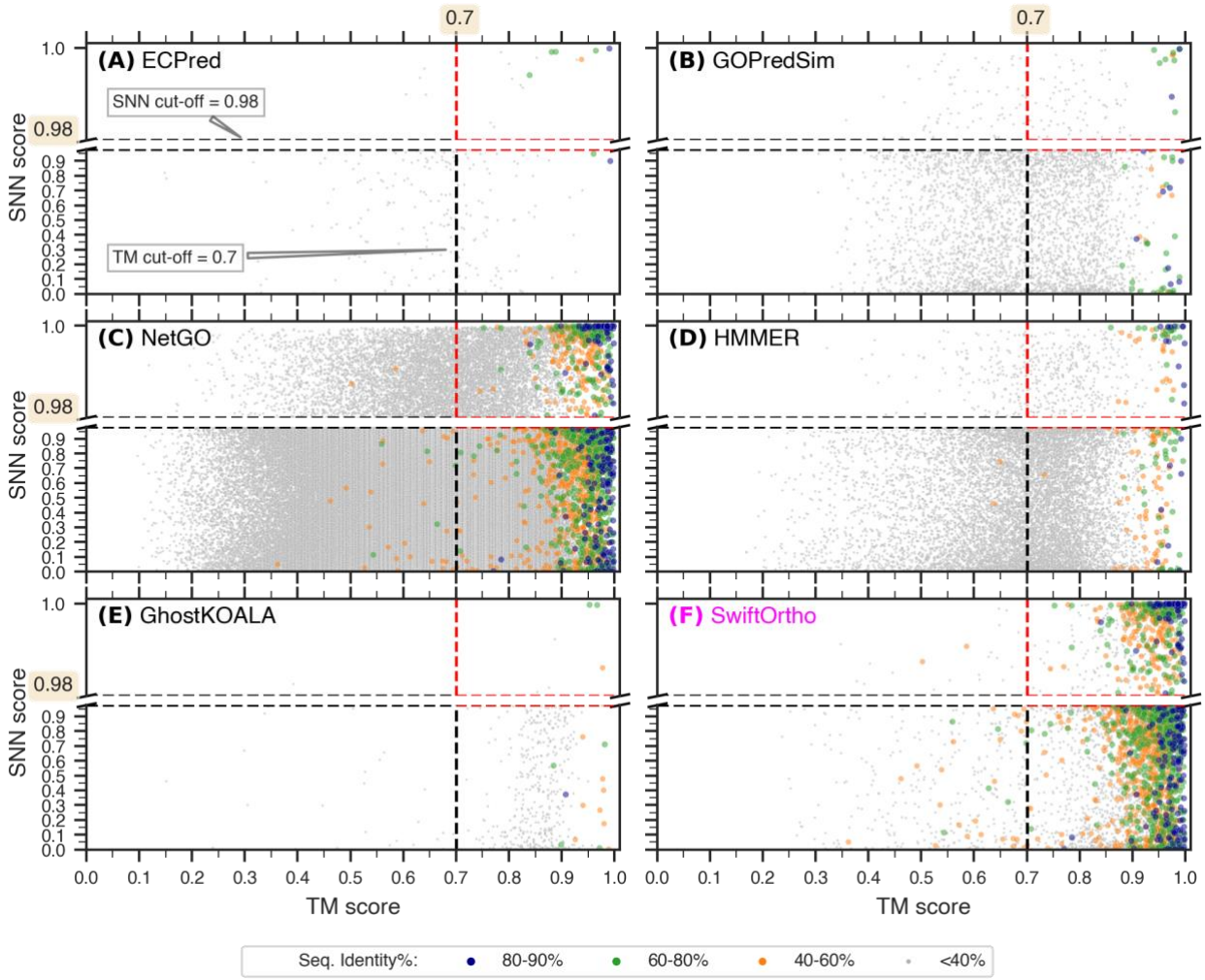

Figure 4: Different annotations capture different functional spaces. Orphan siblings predicted by the top-performing methods **(A-E)** and SwiftOrtho **(F)** occupy different spaces in the TM score (x-axis) vs SNN score (y-axis) space depending on the type annotation (EC, GO, Pfam and orthologs). Each dot in the plot represents a protein-pair and is colored based on the sequence identity. The dashed lines indicate the TM and SNN score cut-offs (0.7 and 0.98) chosen for the defining the orphan sibling set (red dashed lines).

We further analysed our data by binning predicted siblings based on structural (TM-score) and putatively functional similarities (SNN score). Predicted siblings from all methods, except ECPred were enriched in high structurally similar pairs (TM-score =[0.7,1.0]) vs. the low similarity range (TM-score =[0,0.5], **Table 2**). In other words, while most methods capture functional similarity driven by structural similarity, ECpred identified similar enzymatic activity in remotely structurally similar protein pairs as well. Note that ECPred predictions were significantly enriched in the high SNN score space

([0.98,1]). This is not unexpected given that functional convergence is more probable than sequence or even structural convergence[86, 87] and ECPred relies on (sub)sequence and physiochemical feature similarity to predict EC numbers[66].

Different patterns yet were observed in bins reflecting moderate levels of structural similarity (0.5-0.7) and sequence identity (40-60%). GhostKOALA showed a five-fold enrichment in protein pairs with moderate levels of sequence identity and a 14-fold depletion in protein pairs with moderate levels of structural similarity; a similar trend was observed for HMMER. Note that protein pairs with a TM-score over 0.5 are highly likely to share fold-level similarity[83] – a feature that can be expected to reflect function, but does not appear to be useful to the methods reported here.

To summarize, sequence similarity is widely recognized as a key determinant in assessing functional similarity, particularly due to many functional evolution events resulting from gene duplication[88]. However, we hypothesise that existing protein sequence and domain recognition-based methods are biased towards capturing sequence similarity over functional signatures, thus failing to capture analogous evolution, reflect on function diverged between sequence-similar homologs, and identify conserved function among highly diverged siblings[24, 89, 90].

Even considering the incomplete and erroneous functional annotations, everything we currently know about specific proteins and their functions is only a minor fraction of the entire protein universe[91]. However, annotating new proteins based on available data seems to be an inherently flawed proposition. Of the 53 million high-quality predicted ESM structures of microbial proteins extracted from MGnify only 54K (0.1%) had less than 30% sequence identity with UniRef. In turn, the overlap between all of UniProtKB and MGnify is estimated to be less than 1%[46]. That is, quality structure predictions, even with the aid of protein Language Models (pLMs), are limited to known protein families. This could explain the lacklustre inference based on pLM embeddings for orphan proteins (**Figure S1)**. Nevertheless, GoPredSim which leverages function-transfer based on embedding similarity through k-nearest neighbours, is one of our four top performers, underscoring the potential of adapting current deep learning techniques to identify functional similarity among proteins.

Summarizing the findings. With the growing stockpile of sequences, development of accurate functional annotation tools is more essential than ever. A major limitation to the assessment is the lack of large and diverse “ground truth” annotations. In this review, we assess some of the top protein annotation tools on a set of “orphan” proteins. To assess the quality of annotations, we translate the challenge of function prediction into a task of identifying functionally similar protein pairs in the dataset. Careful evaluation across a range of metrics reveals that even the performance of the top methods (ECPred, GOPredSim, HMMER and GhostKOALA) on this data is lacklustre. We note that even though the methods considered herein use different annotation vocabularies, our approach of deriving protein similarity scores enables their comparison.

In this review, we also explored the definition of protein functional similarity in terms of sequence and structure. Machine learning-based models such as ECPred, NetGO and GOPredSim capture more than sequence similarity from input sequences, unlike the more sequence homology-based algorithms. However, functional similarity is not obviously a function of sequence or structural similarity. The robustness of protein conformations paves the way for fully populated this three-dimensional space where diverse or similar sequences fold into diverse or similar structures to carry out the same or different functions as the environment dictates.

Another key observation from our work is that there is a lot of room for improvement in training deep learning models for protein functional annotation. Neither specifically trained methods, nor the cross-domain application of protein embeddings to identify functionally similar pairs showed promising results, highlighting the need for fine-tuning and analysis. While a huge advance has been made in protein structure determination in recent years, similar improvement in function annotation is limited by the lack of ground-truth annotation. However, alternate approaches to evaluation, such the one we put forward, could pave the way for better models.

Data availability

All data used in this study are listed in the main text or deposited in a permanent online data repository. The dataset of orphan proteins and the function similarity scores are available at 10.6084/m9.figshare.c.6737127.

Acknowledgements

We thank Iddo Friedberg (Iowa State University) and Predrag Radivojac (Northeastern University) for their thoughtful discussions on the topic of function. We are grateful to all the developers of methods presented herein for sharing their models and data and for responding to our questions. This work was supported by NAI Grant Number: 80NSSC18M0093. Y.B. was also supported by the NSF (National Science Foundation) CAREER award 1553289.

References

1. Milo, R. What is the total number of protein molecules per cell volume? A call to rethink some published values. *BioEssays* **35**, 1050-1055 (2013).
2. Wiśniewski, J.R., Hein, M.Y., Cox, J. & Mann, M. A “Proteomic Ruler” for Protein Copy Number and Concentration Estimation without Spike-in Standards*. *Molecular & Cellular Proteomics* **13**, 3497-3506 (2014).
3. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
4. Altschul, S.F. & Koonin, E.V. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends Biochem Sci* **23**, 444-447 (1998).
5. Potter, S.C. et al. HMMER web server: 2018 update. *Nucleic Acids Res* **46**, W200-W204 (2018).
6. Eddy, S.R. A new generation of homology search tools based on probabilistic inference. *Genome Inform* **23**, 205-211 (2009).
7. Altschul, S.F., Gertz, E.M., Agarwala, R., Schaffer, A.A. & Yu, Y.K. PSI-BLAST pseudocounts and the minimum description length principle. *Nucleic Acids Res* **37**, 815-824 (2009).
8. Burley, S.K. et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res* **47**, D464-D474 (2019).
9. UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480-D489 (2021).
10. Sayers, E.W. et al. GenBank 2023 update. *Nucleic Acids Res* **51**, D141–D144 (2022).
11. Mistry, J. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412-D419 (2021).
12. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**, 366-368 (2021).
13. Hauser, M., Steinegger, M. & Soding, J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* **32**, 1323-1330 (2016).
14. Zhu, C. et al. Functional sequencing read annotation for high precision microbiome analysis. *Nucleic Acids Res* **46**, e23 (2018).
15. Mahlich, Y., Steinegger, M., Rost, B. & Bromberg, Y. HFSP: high speed homology-driven function annotation of proteins. *Bioinformatics* **34**, i304-i312 (2018).
16. Vanni, C. et al. Unifying the known and unknown microbial coding sequence space. *Elife* **11**, e67667 (2022).
17. Schnoes, A.M., Ream, D.C., Thorman, A.W., Babbitt, P.C. & Friedberg, I. Biases in the Experimental Annotations of Protein Function and Their Effect on Our Understanding of Protein Function Space. *PLOS Computational Biology* **9**, e1003063 (2013).

18. Makrodimitris, S., van Ham, R. & Reinders, M.J.T. Automatic Gene Function Prediction in the 2020's. *Genes (Basel)* **11**, 1264 (2020).

19. Mangul, S., Martin, L.S., Eskin, E. & Blekhman, R. Improving the usability and archival stability of bioinformatics software. *Genome Biol* **20**, 47 (2019).

20. Ramola, R., Friedberg, I. & Radivojac, P. The field of protein function prediction as viewed by different domain scientists. *Bioinform Adv* **2**, vbac057 (2022).

21. Todd, A.E., Orengo, C.A. & Thornton, J.M. Sequence and structural differences between enzyme and nonenzyme homologs. *Structure* **10**, 1435-1451 (2002).

22. Sampaleanu, L.M., Vallee, F., Slingsby, C. & Howell, P.L. Structural studies of duck delta 1 and delta 2 crystallin suggest conformational changes occur during catalysis. *Biochemistry* **40**, 2732-2742 (2001).

23. Gherardini, P.F., Wass, M.N., Helmer-Citterich, M. & Sternberg, M.J. Convergent evolution of enzyme active sites is not a rare phenomenon. *J Mol Biol* **372**, 817-845 (2007).

24. Todd, A.E., Orengo, C.A. & Thornton, J.M. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* **307**, 1113-1143 (2001).

25. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).

26. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

27. Yeats, C. et al. Gene3D: comprehensive structural and functional annotation of genomes. *Nucleic Acids Res* **36**, D414-418 (2008).

28. Saier, M.H. et al. The Transporter Classification Database (TCDB): 2021 update. *Nucleic Acids Res* **49**, D461-D467 (2021).

29. McDonald, A.G. & Tipton, K.F. Enzyme nomenclature and classification: the state of the art. *FEBS J* **290**, 2214-2231 (2021).

30. Tipton, K. & Boyce, S. History of the enzyme nomenclature system. *Bioinformatics* **16**, 34-40 (2000).

31. Tipton, K.F. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement: corrections and additions. *Eur J Biochem* **223**, 1-5 (1994).

32. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* **51**, D587-D592 (2023).

33. Zhou, N. et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol* **20**, 244 (2019).

34. Radivojac, P. et al. A large-scale evaluation of computational protein function prediction. *Nat Methods* **10**, 221-227 (2013).

35. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102-2110 (2022).

36. Elnaggar, A. et al. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. *bioRxiv* (2020).

37. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

38. Lin, Z. et al. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*, 2022.2007.2020.500902 (2022).

39. AlQuraishi, M. Machine learning in protein structure prediction. *Curr Opin Chem Biol* **65**, 1-8 (2021).

40. Fenoy, E., Edera, A.A. & Stegmayer, G. Transfer learning in proteins: evaluating novel protein learned representations for bioinformatics tasks. *Brief Bioinform* **23**, bbac232 (2022).

41. Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T. & Rost, B. Embeddings from deep learning transfer GO annotations beyond homology. *Scientific Reports* **11**, 1160 (2021).

42. Clark, W.T. & Radivojac, P. Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics* **29**, i53-61 (2013).

43. Wang, J.Z., Du, Z., Payattakool, R., Yu, P.S. & Chen, C.F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**, 1274-1281 (2007).

44. Jiang, Y. et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology* **17**, 184 (2016).

45. Zhao, C. & Wang, Z. GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. *Sci Rep* **8**, 15107 (2018).

46. Mitchell, A.L. et al. MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res* **48**, D570-D578 (2020).

47. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722-2728 (2013).

48. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702-710 (2004).

49. Suzek, B.E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926-932 (2015).

50. Mirdita, M. et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res* **45**, D170-D176 (2017).

51. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

52. Mahlich, Y. et al. Rethinking bacterial relationships in light of their molecular abilities. *bioRxiv*, 2022.2011.2028.518265 (2022).

53. van Kempen, M. et al. Foldseek: fast and accurate protein structure search. *bioRxiv* (2022).

54. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-462 (2016).

55. Galperin, M.Y., Kristensen, D.M., Makarova, K.S., Wolf, Y.I. & Koonin, E.V. Microbial genome analysis: the COG approach. *Brief Bioinform* **20**, 1063-1070 (2019).

56. Gligorijevic, V. et al. Structure-based protein function prediction using graph convolutional networks. *Nat Commun* **12**, 3168 (2021).

57. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422-429 (2020).

58. Kabir, A. & Shehu, A. GOProFormer: A Multi-Modal Transformer Method for Gene Ontology Protein Function Prediction. *Biomolecules* **12**, 1709 (2022).

59. Yao, S. et al. NetGO 2.0: improving large-scale protein function prediction with massive sequence, text, domain, family and network information. *Nucleic Acids Res* **49**, W469-W475 (2021).

60. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).

61. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).

62. Bileschi, M.L. et al. Using deep learning to annotate the protein universe. *Nature Biotechnology* **40**, 932-937 (2022).

63. Sanderson, T., Bileschi, M.L., Belanger, D. & Colwell, L.J. ProteInfer, deep neural networks for protein functional inference. *Elife* **12** (2023).

64. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol* **428**, 726-731 (2016).

65. Aramaki, T. et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251-2252 (2020).

66. Dalkiran, A. et al. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics* **19**, 334 (2018).

67. Queiros, P., Novikova, P., Wilmes, P. & May, P. Unification of functional annotation descriptions using text mining. *Biol Chem* **402**, 983-990 (2021).

68. Bepler, T. & Berger, B. Learning protein sequence embeddings using information from structure. *ArXiv* **abs/1902.08661** (2019).

69. Lu, A.X., Zhang, H., Ghassemi, M. & Moses, A. Self-Supervised Contrastive Learning of Protein Representations By Mutual Information Maximization. *bioRxiv*, 2020.2009.2004.283929 (2020).

70. Hoarfrost, A., Aptekmann, A., Farfanuk, G. & Bromberg, Y. Deep learning of a bacterial and archaeal universal language of life enables transfer learning and illuminates microbial dark matter. *Nat Commun* **13**, 2606 (2022).

71. Heinzinger, M. et al. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* **20**, 723 (2019).

72. Dallago, C. et al. Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets. *Curr Protoc* **1**, e113 (2021).

73. Hu, X. & Friedberg, I. SwiftOrtho: A fast, memory-efficient, multiple genome orthology classifier. *Gigascience* **8**, giz118 (2019).

74. Chao, K.-M., Pearson, W.R. & Miller, W. Aligning two sequences within a specified diagonal band. *Bioinformatics* **8**, 481-487 (1992).

75. Dongen, S.V. Graph Clustering Via a Discrete Uncoupling Process. *SIAM Journal on Matrix Analysis and Applications* **30**, 121-141 (2008).

76. Clark, W.T. et al. Assessment of predicted enzymatic activity of alpha-N-acetylglucosaminidase variants of unknown significance for CAGI 2016. *Hum Mutat* **40**, 1519-1529 (2019).

77. Camon, E. et al. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res* **32**, D262-266 (2004).

78. Huntley, R.P. et al. The GOA database: gene Ontology annotation updates for 2015. *Nucleic Acids Res* **43**, D1057-1063 (2015).

79. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261-272 (2020).

80. Vaswani, A. et al. Attention Is All You Need. *Adv Neur In* **30**, 5998-6008 (2017).

81. Hugo, D.-T. et al. The Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics. *bioRxiv*, 2023.2001.2011.523679 (2023).

82. Li, F.-Z., Amini, A.P., Yang, K.K. & Lu, A.X. Pretrained protein language model transfer learning: is the final layer representation what we want? *Machine Learning for Structural Biology Workshop, NeurIPS 2022* (2022).

83. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889-895 (2010).

84. Gabaldon, T. & Koonin, E.V. Functional and evolutionary implications of gene orthology. *Nat Rev Genet* **14**, 360-366 (2013).

85. Nehrt, N.L., Clark, W.T., Radivojac, P. & Hahn, M.W. Testing the ortholog conjecture with comparative functional genomic data from mammals. *PLoS Comput Biol* **7**, e1002073 (2011).

86. Doolittle, R.F. Convergent evolution: the need to be explicit. *Trends Biochem Sci* **19**, 15-18 (1994).

87. Almonacid, D.E., Yera, E.R., Mitchell, J.B. & Babbitt, P.C. Quantitative comparison of catalytic mechanisms and overall reactions in convergently evolved enzymes: implications for classification of enzyme function. *PLoS Comput Biol* **6**, e1000700 (2010).

88. Taylor, J.S. & Raes, J. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet* **38**, 615-643 (2004).

89. Omelchenko, M.V., Galperin, M.Y., Wolf, Y.I. & Koonin, E.V. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biol Direct* **5**, 31 (2010).

90. Galperin, M.Y., Walker, D.R. & Koonin, E.V. Analogous enzymes: independent inventions in enzyme evolution. *Genome Res* **8**, 779-790 (1998).

91. Ghatak, S., King, Z.A., Sastry, A. & Palsson, B.O. The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. *Nucleic Acids Res* **47**, 2446-2454 (2019).

Figure captions

Fig. 1: The limits of protein function annotations.
Of the complete set of proteins (entire circle), containing known/previously observed proteins (blue/black dashed circle outline) and unknown/not-yet seen proteins, some fraction carries out unknown functions (fraction of circle in red) rather than known (fraction in white) ones. Existing experimental (and homology-based) protein function annotations (blue oval) cover a small part of the complete protein sequence space. The stockpile of orphan proteins, i.e. those lacking annotation and having no known homologs, is growing as we explore previously untouched environments with better and faster gene/protein capture tools. This manuscript is focused on evaluating function prediction method performance for this orphan set.

Fig. 2: Defining the Orphan sibling dataset.
Orphan protein pairs with similar functionality are termed "Orphan siblings". In this study, Orphan siblings are identified from a dataset of 309,549 orphan protein pairs using two orthogonal scores: TM score computed from structural alignment and SNN functional similarity score from cDNA sequences. Plot (A) shows the number of orphan pairs considered as siblings at different TM score and SNN score cut-offs. Plot (B) shows the distribution of all 309,549 protein pairs on TM and SNN score space. Each dot in the plot represents a protein-pair and is coloured based on the sequence identity between them: < 40% (gray), 40-60% (orange), 60-80% (green) and 80-90% (darkblue). The dotted red/black lines indicate the TM and SNN score cut-offs (0.7 and 0.98) chosen for the comparison. The red lines border the TM-SNN space of the orphan siblings.

Fig. 3: Comparison of prediction performance in each annotation category.
The bar plots show the highest (a) ∆Similarity score, (b) F1 score and (c) Area under the Precision-Recall curve for the prediction of functionally similar proteins based on predicted protein annotations: Enzyme commission number (EC1 – first digit, EC3 – third digit), Gene Ontology (MF), Pfam domains and Ortholog groups. The highest scores were calculated by varying the prediction score threshold for annotation terms at TM-score and SNN cutoffs of 0.7 and 0.98. The double line separates direct functional annotations: EC and GO annotations from Pfam and Ortholog definitions. The gray dotted horizontal lines indicates the performance of the random baseline classifier.

Fig. 4: Variations in performance over TM-score cut-offs.
The plots (A-D) show the ∆Similarity score (see Methods section for details) for the prediction of functionally similar proteins based on predicted protein annotations: Enzyme commission number, Gene Ontology (MF), Pfam and Ortholog groups, respectively. Similarly, plots (E-H) display the F1max for the four groups of annotations. The highest ∆Similarity and $F1_{max}$ scores were computed at SNN cutoff of 0.98 and at different TM-score cutoffs [0, 0.98] by varying the prediction score threshold for annotation terms. The average performance of 100 random baseline classifiers is plotted for comparison in each panel (dotted black line). The bottom and top horizontal axes indicate the TM score cut-off and the corresponding number of positive protein pairs – orphan siblings in the dataset.

Fig. 5: Different annotations capture different functional spaces.
The plots (A-E) show orphan siblings predicted by the top-performing methods. The protein pairs predicted as orphan siblings occupy different spaces depending on the type annotation: EC, GO, Pfam and orthologs. Each dot in the plot represents a protein-pair and is coloured based on the sequence identity between them: < 40% (gray), 40-60% (orange), 60-80% (green) and 80-90% (darkblue). The dotted red/black lines indicate the TM and SNN score cut-offs (0.7 and 0.98) chosen for the comparison. The red lines border the TM-SNN space of the orphan siblings. The orthologs pairs identified by SwiftOrtho are plotted in panel F for comparison.

Tables

Table 1: Protein function prediction performance

Tool	Term	ΔS	$F1_{max}$	PR AUC	ROC AUC	Precision	Recall	Recall (90)	Number of predicted positives[a]
ECPred	EC1	**0.16**	**0.68**	**0.72**	**0.59**	**0.56**	**0.88**	**0.90**	**33**
ECPred	EC3	**0.26**	**0.69**	**0.70**	**0.67**	**0.65**	**0.75**	**1.00**	**9,122**
Mantis	EC1	0.04	0.66	0.71	0.52	0.51	0.91	0.91	692
Mantis	EC3	0.04	0.62	0.64	0.53	0.52	0.77	1.00	780
DeepFRI	GO	0.10	0.65	0.59	0.60	0.56	0.77	0.85	10,118
DeepGOPlus	GO	0.01	0.31	0.40	0.54	0.64	0.20	1.00	2,256
GOPredSim	GO	**0.16**	**0.68**	**0.66**	**0.61**	**0.59**	**0.81**	**0.81**	**129**
Mantis	GO	0.11	0.64	0.60	0.58	0.57	0.72	1.00	1,120
NetGO	GO	0.09	**0.67**	0.67	0.68	0.57	0.83	0.90	10,468
ProteInfer	GO	0.03	0.46	0.44	0.56	0.60	0.37	1.00	3,476
GhostKOALA	KO	**0.09**	**0.66**	**0.70**	**0.55**	**0.53**	**0.86**	**0.86**	**11**
Mantis	COG	0.05	0.62	0.65	0.53	0.52	0.77	1.00	2496
HMMER	Pfam	**0.23**	**0.70**	**0.76**	**0.66**	**0.54**	**0.96**	**0.90**	**3,638**
Mantis	Pfam	0.18	0.63	0.65	0.64	0.58	0.70	1.00	3,346
ProtCNN	Pfam	0.00	0.13	0.42	0.52	0.78	0.07	1.00	12,438

Performance of the selected, top-performing methods is identified in bold.
[a] The number of orphan pairs predicted as positives by each method at a Recall of 90% is reported.

Table 2: Characteristics of protein pairs predicted as siblings

	Range	Total Orphan pairs	Siblings	ECPred	GOPredSim	NetGO	HMMER	GhostKOALA	SwiftOrtho
Sequence Identity	<40%	307,434 (99.3%)	5,576^{--} (89.7%)	219^{--}	3,724^{--}	200,223^{--}	10,696^{--}	456^{--}	2,150^{--}
	40-60%	1,155 (0.4%)	314^{++} (5.0%)	1	11	1,131^{++}	88^{++}	9^{++}	1,019^{++}
	60-80%	720 (0.2%)	234^{++} (3.7%)	5^{++}	32^{++}	714^{++}	58^{++}	6^{++}	717^{++}
	80-90%	240 (0.1%)	95^{++} (1.5%)	2^{+}	10^{+}	237^{++}	19^{+}	1	239+
TM-score	0-0.5	39,675 (14.5%)	0^{--}	22^{-}	301^{--}	25,666^{--}	891^{--}	10^{--}	118+
	0.5-0.7	133,206 (45.0%)	0^{--}	111	1,547^{--}	88,625^{--}	4,039^{--}	15^{--}	249^{--}
	0.7-1.0	136,668 (40.4%)	6,219^{++}	94	1,929^{++}	88,014^{++}	5,931^{++}	447^{++}	3,758^{++}
SNN score	0-0.5	102,620 (50.1%)	0^{--}	88^{--}	1,970^{+}	98,533^{--}	5,726^{++}	298^{++}	1,436+
	0.5-0.98	181,495 (45.8%)	0^{--}	120^{+}	1,644^{-}	93,997^{++}	4,584^{--}	167^{--}	1,841
	0.98-1.0	25,434 (4.1%)	6,219^{++}	19^{+}	163	9,775^{++}	551^{++}	7^{-}	848^{++}

Values indicate the number of siblings predicted by each method in the given range of sequence identity, TM-score, or SNN score. The values were compared against the corresponding counts in the entire dataset of orphan pairs via the two-sided Fischer's exact test. +/- denotes a significant increase/decrease with p-value in [0.001, 0.05] range, while ++/-- denotes p-value of <0.001